J.T. Yao, V.V. Raghvan, G.Y. Wang (Eds.): WSS'04, pp. 142-148, 2004.

MICE-*f*: Financial Reviews Analysis using Category Model

Saravadee Sae Tan[1], Gan Keng Hoon[1], Tang Enya Kong[1], Chan Huah Yong[2]

[1]*Computer Aided Translation Unit (UTMK)* [2]*Grid Computing Lab*

School of Computer Sciences
Universiti Sains Malaysia
11800 Penang Malaysia
{saratan|khgan|enyakong|hychan}@cs.usm.my

Abstract

In the domain of financial, financial news, articles, reports about financial reviews are helpful and important information which give investors or financial analyst an indication to help decision making in financial matters. However, due to the volume of this information and the diversity of financial topics, it is difficult for a human to track and interpret each of them in a consistent manner.

Based on this motivation, we propose that this information can be classified into categories, in which the categories are based on a particular objective and goal as required by a user. In each category, the financial information is further classified based on a status indicator, to reflect the positive, neutral or negative status in term of financial outlook stated in the information. Thus, user can directly focus on interested financial topic, and get an indication about the financial outlook of that topic.

For classification purpose, we propose to combine linguistic technique and statistical technique to select features to represent the categories and status indicators.

In this work, we present the architecture of MICE-f that will crawl financial information from multiple financial sources and classify them based on the defined categories and status indicators.

1. Introduction

In the domain of financial, large quantities of articles, news, reports about financial reviews are generated daily. With the technology of Internet, this information can be easily retrieved using online sources such as Reuters, Bloomberg, CNN Financial Network etc. This information contains a wealth of financial knowledge that can be used to help decision making in financial matters. Human interpretation (e.g. financial analyst) is needed to analyze and transform the textual information into useful knowledge, in order to get a summary or conclusion about the positive or negative status of financial outlook reflected by the financial reviews. However, due to the tremendous amounts of information, it is impossible for a user to go through every single piece of information from various sources every day.

Although financial information had been classified by various financial sources into categories such as economy, politic, market etc, we have different users with different needs and objectives when accessing this information. Therefore, we would like to orientate the retrieval and classification of financial information based on the interest of each user.

In this paper, we propose a classification method that is able to classify the financial information based on categories defined by user. Intuitively speaking, the user has the flexibility to define the concepts in which each category represents. For this reason, he may foresee the information or content being classified in each category thus can directly focus on information he is seeking. In each category, the information is further classified based on three status indicators i.e. positive, neutral and negative to reflect the financial outlook status of the information.

As an extension to our previous work in text classification for general web information [6], we propose a system, MICE-*f* that is orientated to the financial domain. In MICE-*f*, we

i) Justify our Category Model (CM) to represent categories related to financial domain. The Category Model consists of two levels. The first level is to organize financial information into pre-defined categories. The second level is to classify the financial information based on our status indicators, whether a piece of financial information reflect positive, neutral or negative financial outlook with respect to the category.
ii) Enhance our feature selection technique [7] by including linguistic analysis. Due to differences of financial text (more precise and specific) compared to general web text, we have identified two types of features to be selected for our classification purpose,

i.e. *concept features* which represent the categories and *descriptive features* which represent status indicators.

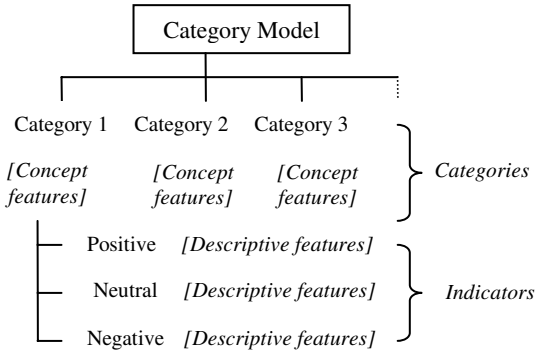


Figure 1: The category model for MICE-*f*.

In MICE-*f*, we adapt the Category Model in Figure 1 as the basis to classify financial information. MICE-*f* allows crawling of desired financial information such as news, reports, commentaries, articles from identified financial sources, classifying them based on user-defined categories as well as the status indicators. The overall idea is to assist users to effectively focus and attend to financial information based on categories relevant to their needs rather than navigating through a pool of overwhelming financial information.

2. The Methodology

2.1 The Nature of Financial Text

Financial texts (e.g. financial news, financial reports, commentaries etc) are usually more compact and straight forward compared to other natural language text like stories, web pages, electronic mail etc. Most financial texts have specific and objective contents. We can easily recognize a particular event (e.g. company takeover, company merge, management succession, election etc) from the text as well as infer whether the information reflects positive or negative status in term of financial outlook [10]. For example,

Drugs giants merge
UK drugs giants Glaxo Wellcome and SmithKline Beecham have confirmed their plans to merge into the world's biggest pharmaceuticals group…
…….
The shares had risen sharply on Friday when the merger talks were confirmed …

From the financial text, we can easily recognize that the news is about company merge and can infer that this news reflect positive status of financial outlook.

Based on the nature of financial text, we propose a classification model which allow user to directly focus on a particular financial topic, such as "company merge", and followed by indication on whether the text is good or bad.

2.2 The Concept of Category Model

Our Category Model has two levels (as in Figure 2). The first level consists of a set of categories defined by user. The second level consists of three indicators which are positive, neutral and negative.

For example, at the first level, user can define a category called 'Company Activities' that represent all events related to company changes, such as company takeover, company merge, management succession, recruitment, and etc. In a more specific scenario, user can define categories "Election", "Governance Transparency" that reflect "Political" issues in a country. The categories should help to focus on certain financial analysis objective.

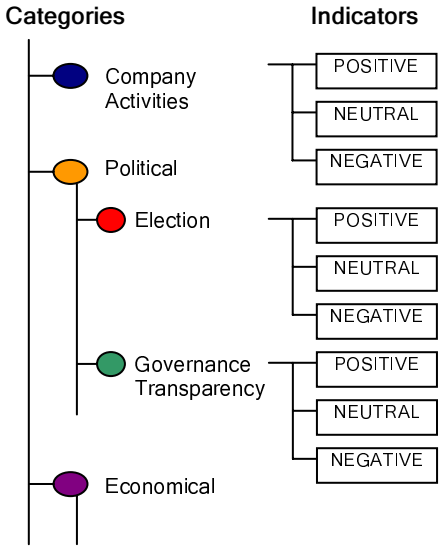


Figure 2: An example set of categories.

At the second level, there are three status indicators reflecting the financial outlook of the information for the defined category. Considering the nature of financial information, the following three status indicators are considered to be pertinent:

i) Positive – information which show good evidence of financial outlook.
ii) Neutral – information which did not mention anything about financial well-being or the influence to financial outlook is unclear.

iii) Negative information which show bad evidence of financial outlook.

In the Category Model, the concept of a category or a status indicator is reflected by a set of characteristic keywords (also known as features). In this work, we identified two types of features to be selected for our model.

i) *concept feature* that can express the concepts and contents of a category.
ii) *descriptive feature* to reflect positive, negative or neutral financial outlook of the information in a status indicator.

2.3 Training the Category Model

The selection of features into the Category Model is the crucial part in our work, as the features selected directly represent the meaning and concept of the defined categories and their status indicators. For each *category-indicator* pair, a number of corresponding financial texts have to be prepared for training purpose (refer to Figure 3). These financial texts should reasonably reflect the concept of the category and financial outlook status they belong.

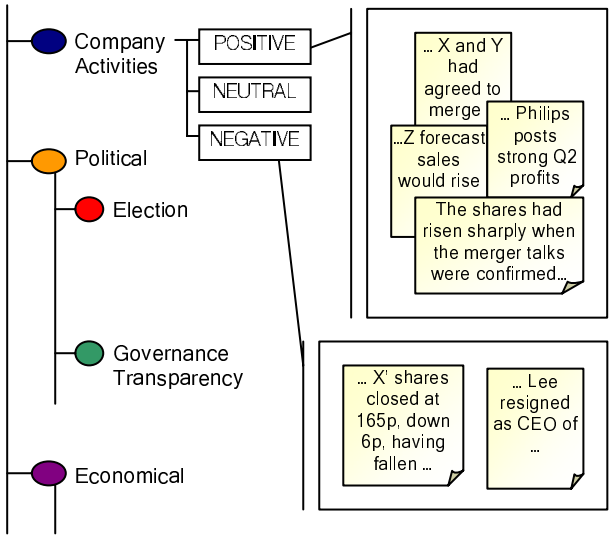


Figure 3: An example financial training texts for Company Activities .

Two main tasks in training a Category Model are:

i) Identify and generate candidate features from the training texts. Candidate features are potential keywords to be selected as *concept features* and *descriptive features*.
ii) From the candidate features, select an optimum set of keywords as *concept features* and *descriptive features*.

2.3.1 Text Parsing

In this paper, we propose to use Partial Parsing technique to analyze a financial text in order to extract relevant information to be considered as candidate features. Information such as subject of a sentence, object of a sentence, noun phrase and etc can be easily identified by analyzing the syntactic structure of a sentence.

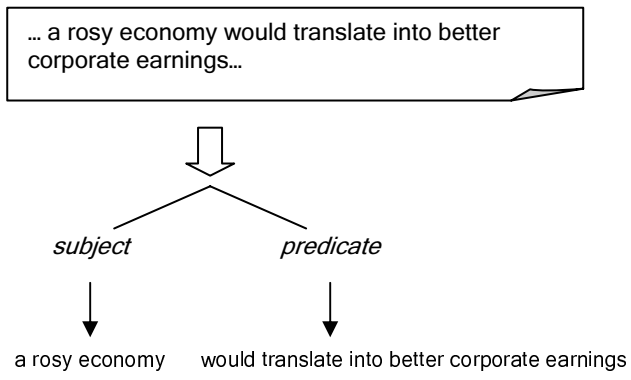
Partial Parsing is a linguistic technique to analyze syntactic structure of natural language texts. Partial Parsing perform partial analysis of the syntactic structures in a text. There are several possible levels of partial parsing: from identification of base noun phrases, to identification of chunks and to identification of clauses [2] [5] [9].

Clause Identification is a method in Partial Parsing to identify clauses in a text. A clause is a sequence of words in a sentence that contains a subject and a predicate [2]. Since the financial text selected for training purpose reflect the concepts of its category, we can make *assumptions* that:

i) The subject of a clause in a financial text is assumed to be related to the concepts or contents of the category the financial text belongs. Thus, the subject can be further analyzed to extract relevant word/phrase to be considered as a candidate for our *concept features*.
ii) The predicate of a clause may describe the action of the subject or may contain information about the influence of an event mention in the subject. This action or influence may have a direct relation with the financial outlook of the event. Thus it can be further analyzed to extract relevant word/phrase to be considered as a candidate for our *descriptive features*.

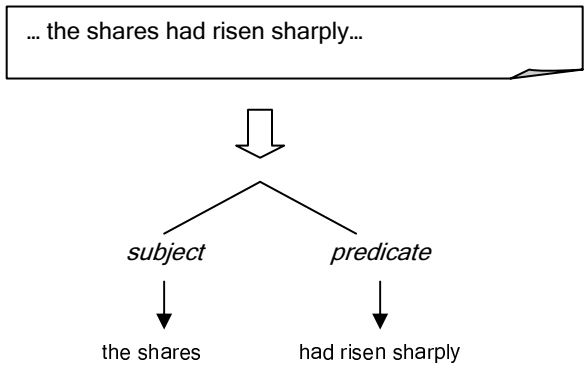
Here are examples of clauses obtained from financial news.

Example 1:



In this example clause, the event is *rosy economy* and this event has caused a *better corporate earnings*. Thus, *'rosy economy'* can be considered as a candidate for *concept feature* and *'better corporate earnings'* can be considered as a candidate for *descriptive feature*.

Example 2



In this clause, the subject *'shares'* can be considered as a candidate for *concept feature* and the behavior of the subject, *'risen sharply'* can be considered as a candidate for *descriptive feature*.

The methodology of Text Parsing is not finalized and further research will be carried out on it.

2.3.2 Feature Selection

In our methodology, Feature Selection technique is applied to select the *concept features* and *descriptive features* from the candidate features in order to compose the Category Model.

Feature Selection is a process that chooses a subset of features from the original set of features so that the dimensionality of feature space is optimally reduced according to a certain criterion. This tends to produce classification models that are simpler, clearer and computationally less expensive [4].

Various approaches of feature selection have been developed for dimensionality reduction in a classification task. Basically, these methods can be broadly divided into 2 main approaches, (i) Feature Selection in Machine Learning, and (ii) Feature Selection in Text Learning. Feature Selection in Machine Learning traverse a feature space and evaluate every candidate feature subset in order to find the best subset. These methods are less practical when the number of features is large. On the other hand, Feature Selection in Text Learning evaluates every feature independently, in which a scoring criterion is used to measure the goodness of a feature. All features are sorted in a list and a predefined number of best features are selected. However, the number of features to be selected is a main experimental issue in these methods [4][8].

Feature selection approach adopted in this paper is from the author's previous work. It combines the idea from both methods of feature selection in machine learning and text learning [7] [8]. All features are sorted in a list using a feature weighting function. An optimum set of features is selected by finding a cut-off point in the list using a consistency measure.

Feature Weighting

Statistical information of a feature, i.e. *frequency distribution of the feature across categories,* is used to indicate the importance of a feature in term of the discriminating power between categories. This comes from two major concerns. A feature is considered as representative if it appears many times *within* a text. On the other hand, a feature is regarded as not informative if it appears too many times *among* texts [1]. In our weighting function, these two aspects are taken as the basis in weighting a feature [7]:

i) ***Feature Frequency*** of a feature denotes the frequency occurrences of the feature in a category. The rational behind is that the ability of a feature in discriminating categories depends on how frequent it occurs in a category as against the other categories
ii) ***Document Frequency*** of a feature denotes the number of documents/texts in a category in which the feature occurs at least once. The main idea is that features that occur in more documents in a category against other categories are more discriminative than features that occur in many documents in many categories.

Every candidate feature is assigned a score using the Feature Weighting function. The score can reflect the significance of a feature in term of the discriminating power between categories. All candidate features are sorted from the most significant to the least significant, and top-*N* features are selected to compose the Category Model.

Consistency Measure

The size of Category Model is a main concern in the processing speed of our classification algorithm. Thus, it is important to control the set of features selected, (the value of *N*) in order to compose an optimum Category Model. We expect that the selected set of features are informative enough to represent the concepts of the categories, neither too few to miss the semantics or too many to burden the processing speed.

In our approach, a selected *feature subset* is evaluated by *class separability* measure. The feature subset is

considered 'optimal' when it maximizes the class separability within a corpus (a collection of training texts). Consistency measure is a conservative way of achieving class separability. It does not attempt to maximize the class separability but tries to retain the power of class separability defined by the original set of features. The idea is to find the smallest set of features that can distinguish the user defined categories as well as the full set of the candidate features [3].

2.3.3 Category Model

Concept feature and *descriptive feature* selected by Text Parsing and Feature Selection are represented in our Category Model. Every category has a set of *concept feature* to differentiate it from other categories. Similarly, every status indicator of the category is represented by a set of *descriptive feature*.

A simple and frequently used representation is the feature vector representation. In our representation, each category or status indicator is characterized by a Boolean vector. All vectors are embedded in a *feature space* where each dimension corresponds to a feature (*concept feature* or *descriptive feature*). In a Boolean vector, each feature has a Boolean value that indicates whether the feature appears or not. The Category Model representation is visualized in Figure 4.

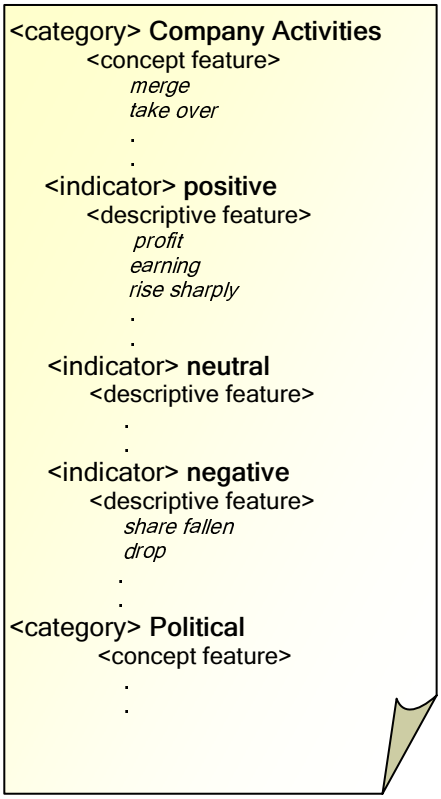


Figure 4: An example of Category Model

3. The Architecture of MICE-*f*

The architectural design of MICE-*f* consists of three major components, i.e. Category Model Generator, Information Crawler and Information Classifier. Upon receiving a request from user, MICE-*f* submits the query to user specified information sources, retrieves the financial information, and processes the information by classifying them into appropriate categories as well as indicating the financial outlook status of the information.

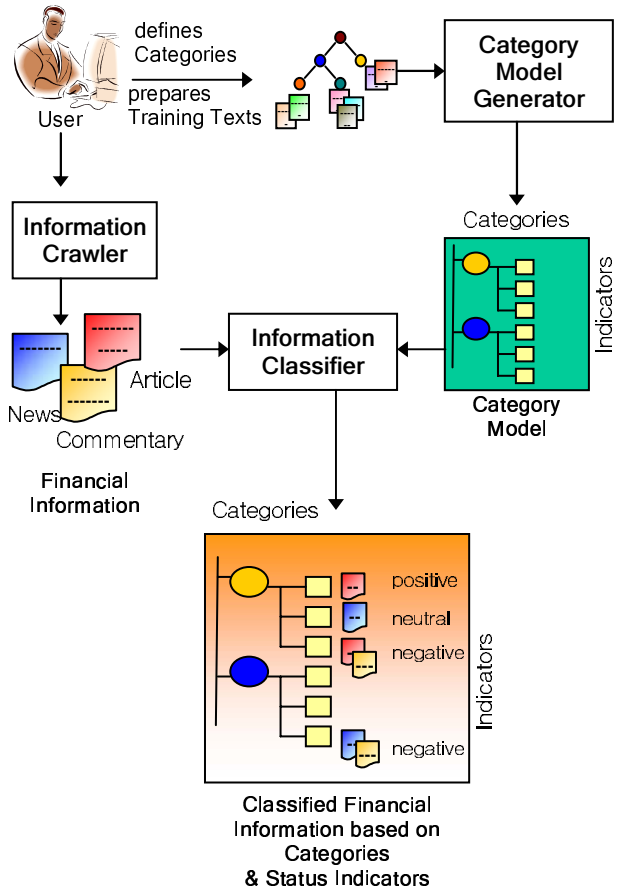


Figure 5: The Architecture of MICE-*f*.

.

3.1 Category Model Generator

The role of Category Model Generator is to learn the concept and characteristics of categories defined by user, and represent them in a Category Model. First, the user has to define a set of categories. The categories should be "well-separated" so that their intersection and overlapping is minimized. Each defined category will be associated to three types of status indicators, i.e. positive, neutral and negative. For each *category-indicator* pair, a set of

financial training texts need to be prepared. The Text Parser (refer to section 2.3.1) will use linguistic technique to analyze the syntactic structure of sentences in the financial texts and extract relevant word or phrase to be considered as candidates of *concept features* and *descriptive features*. From these candidate features, the Feature Selector (refer to section 2.3.2) then uses statistic technique to measure the significance of each candidate feature in term of discriminating power between categories. Finally an optimal set of *concept features* and *descriptive features* are selected to compose the Category Model.

The detail process flow of Category Model Generator is shown in Figure 6.

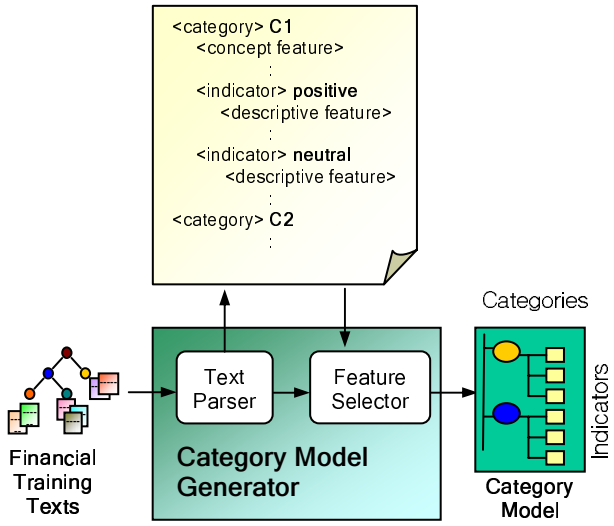


Figure 6: The process of Category Model Generator.

3.2 Information Crawler

There are endless sources for financial information on the World Wide Web. Common sources for financial information are like financial news portal (Google Business News, Yahoo Financial News), company's web sites, and news sites (Reuters, CNN Financial Network, theStar Business, Bloomberg). In the Information Crawler, we can crawl and extract required financial information from multiple sites simultaneously. As the financial information needed by users are varies, this component allows user to specify their required financial sites, company sites or news sites.

When requested by user, the Query Formulator will formulate queries based on the format of the selected financial sources. The Query Dispatcher then sends the queries to these sources simultaneously. From the raw financial texts gathered from these financial sources, Information Extractor will then process and extract useful financial information from the raw financial texts. The process flow for Information Crawler is shown in Figure 7.

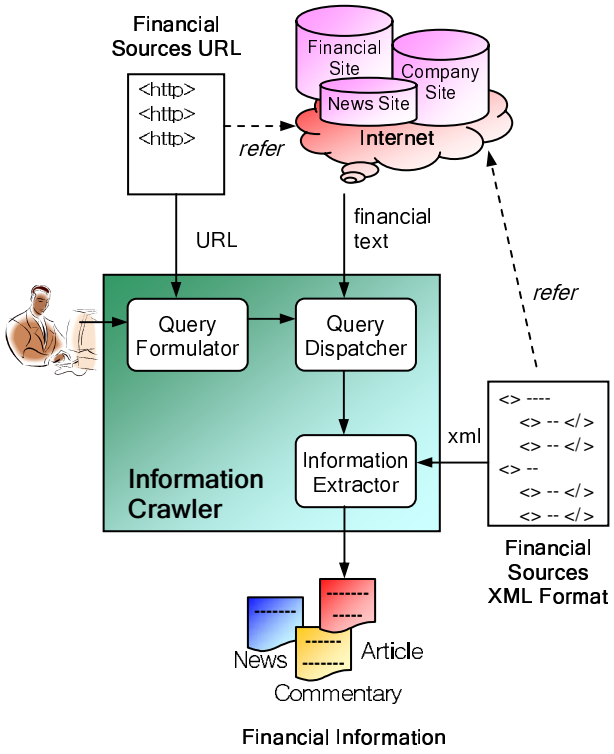


Figure 7: The process of Information Crawler.

3.3 Information Classifier

For each piece of financial information, Information Classifier will analyze its content in order to classify the information into appropriate categories and also infer the status of financial outlook stated in the content.

Text Parser (refer to section 2.3.1) will analyze the syntactic structure of the content and retrieve *concept features* and *descriptive features* found in the content. The *concept features* and *descriptive features* are represented by a vector model with respect to the Category Model vector space (refer to section 2.3.3).

The Similarity Calculator first compares the vector representations for concept features between the financial information and each category in the Category Model. The financial information is assigned to the most similar category. Next, the Similarity Calculator will measure the similarity of the vector representation for descriptive features between the financial information and each status

indicator. An appropriate financial status is assigned to the financial information.

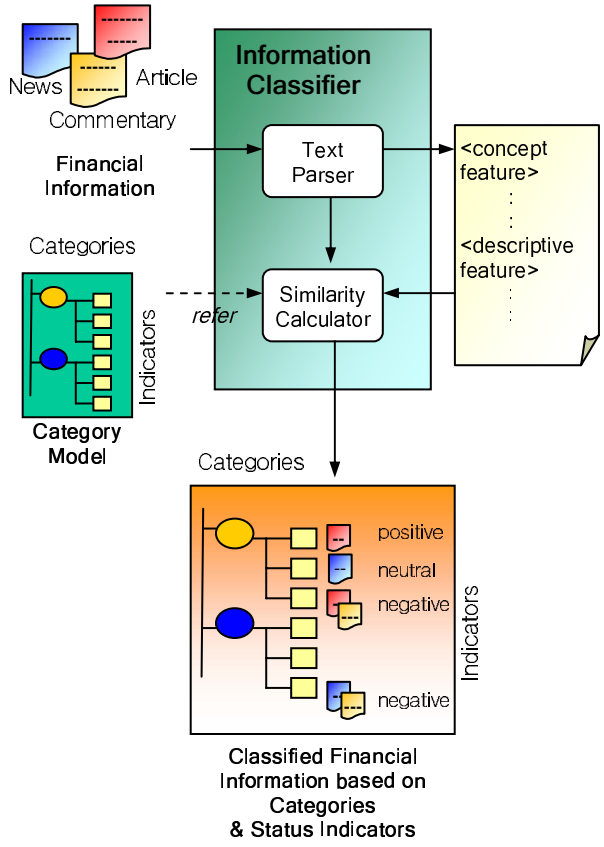


Figure 8: The process of Information Classifier.

4. Conclusion

The increasing number of financial information on the Internet demands a personalized and specialized service to effectively gather and manage the information. We propose an architecture MICE-*f* to crawl financial information from various sources, classify them based on user defined categories and infer the financial outlook status reflect by the information.

Our classification approach combines linguistic technique and statistical technique to select features to represent categories and status indicators. The linguistic technique is still an on-going research and we plan to look into other techniques of parsing in order to extract more specific and accurate information from a text to be considered as *concept feature* and *descriptive feature*.

5. References

[1] C. Liu, "A Survey: Automatic Text Categorization". *CS412 Report*, University of Illinois at Urbana-Champaign, 2004.

[2] E.F. Tjong Kim Sang and H. Dejean, "Introduction to the CoNLL-2001 shared task: Clause identification", In W. Daelemans, and R.Zajac, editors, *Proceedings of CoNLL-2001*, Toulouse, France, 2001, pp53-37.

[3] H. Liu, H. Dash and H. Motoda, "Consistency Based Feature Selection", *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2000),* 2000, pp98-109

[4] M. Sahami and D. Koller, "Toward Optimal Feature Selection", *Proceedings of the 13th International Conference on Machine Learning (ICML 1996)*, San Fransicisco CA, Morgan Kaufmann, pp284-292.

[5] S.P. Abney, "Parsing by chunks", In R.C. Berwick, S.P. Abney, and C. Tenny, editors, *Principle-based parsing: Computation and Psycholinguistics*, Kluwer, Dordrecht, 1991, pp257-278.

[6] S.S. Tan, K.H. Gan, E.K. Tang, S.L. Cheong, S.L. Chan and W.Y. Foo "MICE: Aggregating and Classifying Meta Search Results into Self-Customized Categories", *Proceedings of Web Intelligence (WI 2004)*, Beijing, 2004, accepted as demo-track paper.

[7] S.S. Tan, K.H. Gan, H.Y. Chan, E.K. Tang and S.L. Cheong, "Mapping Search Results into Self-Customized Category Model", *Proceedings of International Conference on Intelligent Information Processing (ICIIP 2004),* Kluwer Academic Publisher, Beijing, October 2004, accepted to appear.

[8] S.S. Tan, "*Topic Hierarchy Annotation using Feature Selection Technique*", MSc Thesis, School of Computer Sciences, Universiti Sains Malaysia, 2002.

[9] X. Carreras, L. Màrquez, V. Punyakanok and D. Roth, "Learning and Inference for Clause Identification", *Proceedings of the 13th European Conference on Machine Learning (ECML 2002)*, Helsinki, Finland, 2002, pp35-47.

[10] Y. Seo, J.A. Giampapa and K. Sycara, *"Text Classification for Intelligent Portfolio Management"*, Technical Report, CMU-RI-TR-02-14, Robotics Institute, Carnegie Mellon University, May 2002.